تمرین دوم

نام و نام خانوادگی: آرشام مسعودی

مقدمه

در این تمرین، هدف ما پیش‌بینی مقدار حق بیمه تعلق گرفته به هر شخص با استفاده از داده‌های موجود در لینک روبه‌رو است: Regression with an Insurance Dataset

این داده‌ها به‌طور خودکار به دو بخش آموزش و تست تقسیم شده‌اند و برای تحلیل و پیش‌پردازش آماده‌سازی شده‌اند؛ این داده‌ها شامل انواع مختلفی از ویژگی‌ها هستند که به دسته‌های زیر تقسیم می‌شوند:

حوزه جمعیت‌شناسی: شامل ویژگی‌هایی مانند سن، جنسیت، وضعیت تاهل و سایر عوامل مرتبط.

حوزه سلامت: شامل اطلاعاتی نظیر سابقه مصرف سیگار، فراوانی ورزش و سایر عوامل بهداشتی.

حوزه جزئیات بیمه‌نامه: شامل نوع بیمه‌نامه، نوع ملک و سایر جزئیات مرتبط با بیمه.

حوزه جغرافیایی: شامل اطلاعات مربوط به محل سکونت افراد و ویژگی‌های جغرافیایی.

مقدار بیمه: که هدف اصلی پیش‌بینی ما می‌باشد.

برای شروع، ابتدا به بررسی ویژگی‌های داده‌ها و شناسایی الگوهای موجود در آن‌ها خواهیم پرداخت. سپس با استفاده از تکنیک‌های مختلف یادگیری ماشین، مدل‌های پیش‌بینی را ایجاد و ارزیابی خواهیم کرد. این فرآیند شامل مراحل زیر است:

۱) **تحلیل داده‌ها:** بررسی و تحلیل ویژگی‌های موجود در داده‌ها برای درک بهتر از ساختار و الگوهای آن‌ها.

۲) **پیش‌پردازش داده‌ها**: شامل پاک‌سازی داده‌ها، مدیریت مقادیر گمشده و نرمال‌سازی ویژگی‌ها به منظور بهبود عملکرد مدل.

۳) **ایجاد مدل:** انتخاب و پیاده‌سازی الگوریتم‌های مناسب برای پیش‌بینی حق بیمه.

۴) **ارزیابی مدل:** بررسی دقت و کارایی مدل‌های ایجاد شده با استفاده از داده‌های تست.

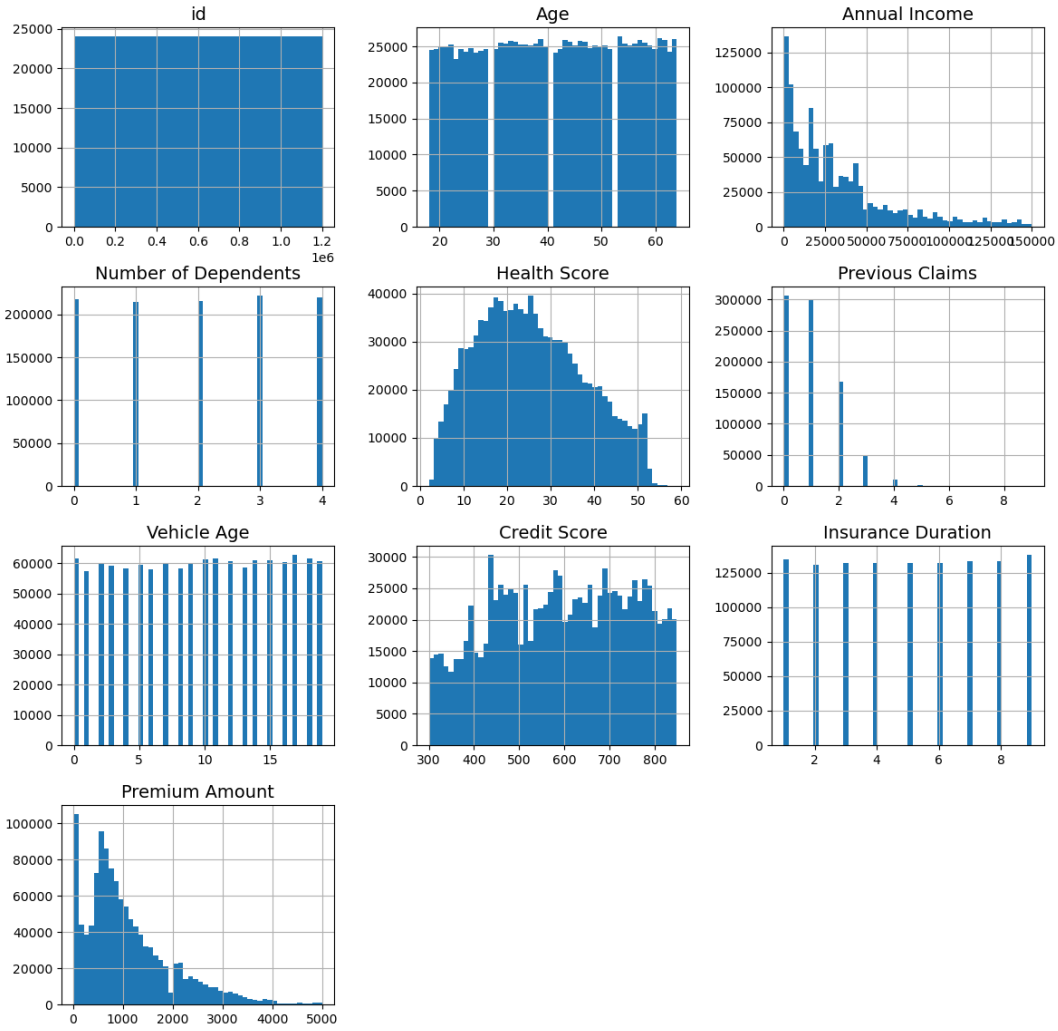
با این رویکرد، می‌توانیم به پیش‌بینی دقیقی از حق بیمه هر شخص دست یابیم.

اکتشاف داده‌ها

برای اکتشاف در داده‌ها، یکی از مهم‌ترین اقدام‌ها رسم هیستوگرام (Histogram) برای هر ویژگی عددی است. این هیستوگرام‌ها به ما کمک می‌کنند تا ببینیم هر ویژگی در چه بازه‌هایی تعداد بیشتری از مقادیر را در دیتاست ما دارد. با تحلیل این هیستوگرام‌ها، می‌توانیم الگوهای موجود در داده‌ها را شناسایی کرده و درک بهتری از توزیع ویژگی‌ها به دست آوریم. نتایج این تحلیل را می‌توانید در تصویر صفحه بعد مشاهده کنید.

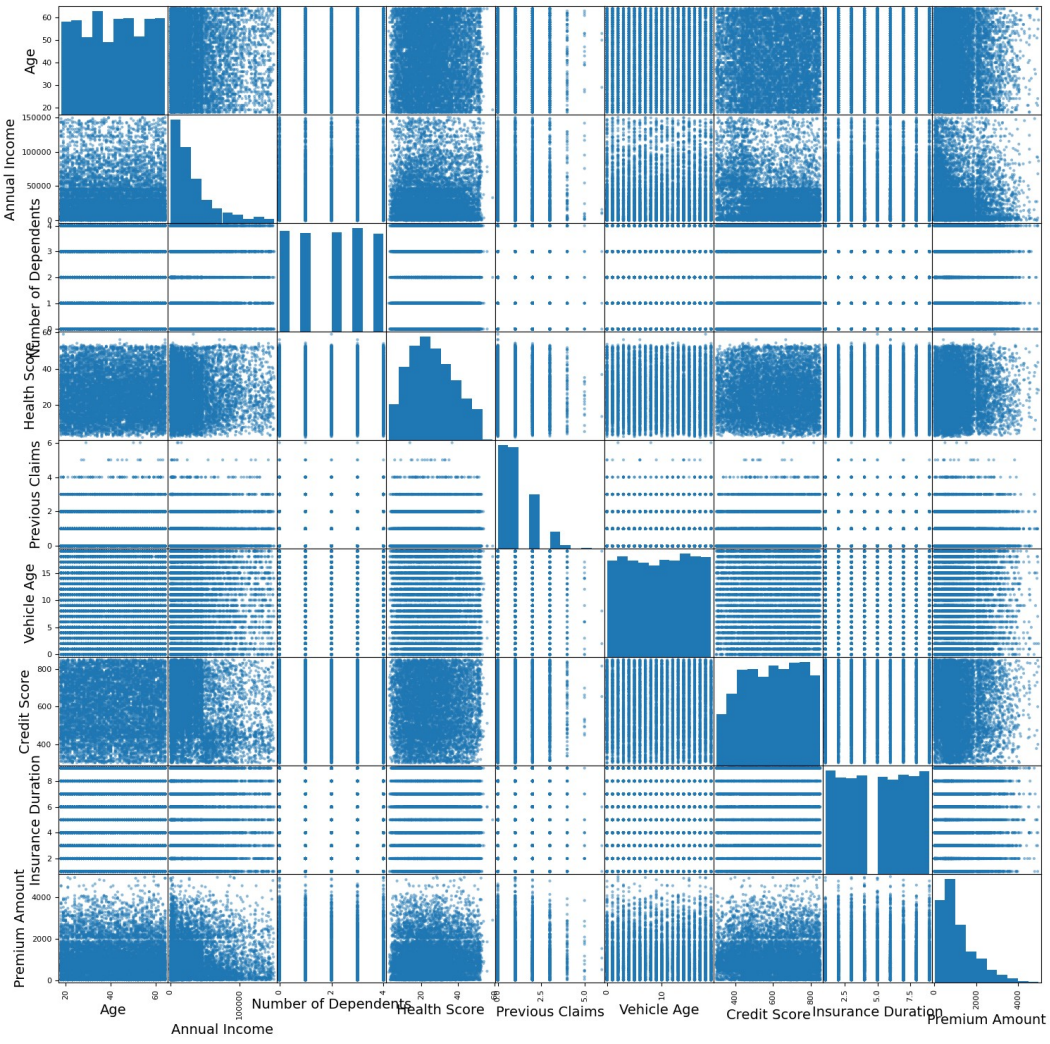
با توجه به توزیع داده‌ها، به نظر می‌رسد که دو صفت Annual Income و Premium Amount دارای چولگی هستند. این موضوع می‌تواند بر کیفیت یادگیری تأثیرگذار باشد. به همین دلیل، در مراحل بعدی، بر روی مقادیر این دو ویژگی لگاریتم اعمال خواهیم کرد تا چولگی آن‌ها از بین برود.

البته باید توجه داشته باشیم که در نهایت، برای بازگرداندن پیش‌بینی‌ها به مقادیر واقعی، باید از تابع وارون لگاریتم، به عبارتی همان تابع نمایی، استفاده کنیم. این کار به ما کمک می‌کند تا نتایج پیش‌بینی شده را به شکل قابل فهم و واقعی ارائه دهیم.



تصویر ۱: هیستوگرام‌های مربوط به صفات عددی

در ادامه، برای کشف ارتباط ویژگی‌ها با یکدیگر، ماتریس scatter را به نمایش آورده‌ایم که در تصویر زیر قابل مشاهده است. با توجه به نمودارهای به‌دست‌آمده، ارتباط واضح و مشخصی بین ویژگی‌ها قابل رؤیت نمی‌باشد.



تصویر ۲: ارتباطات ویژگی‌ها با یکدیگر به صورت شهودی

بنابراین، برای تحلیل بهتر و دقیق‌تر، به یک ویژگی عددی که نشان‌دهنده ارتباطات باشد، یعنی کوواریانس، روی می‌آوریم. این روش به ما کمک می‌کند تا درک بهتری از روابط بین ویژگی‌ها و تأثیر آن‌ها بر یکدیگر به دست آوریم.

	Age	Annual Income	Number of Dependents	Health Score	Previous Claims	Vehicle Age	Credit Score	Insurance Duration	Premium Amount
Age	1.000000	-0.000009	0.001475	0.000881	0.001869	-0.002455	0.002842	-0.000064	-0.002430
Annual Income	-0.000009	1.000000	0.002135	0.025530	0.043065	-0.000500	-0.201423	0.000379	-0.012390
Number of Dependents	0.001475	0.002135	1.000000	0.005152	-0.004123	0.001232	-0.001714	-0.000265	-0.000976
Health Score	0.000881	0.025530	0.005152	1.000000	0.001989	0.000316	0.012016	0.002487	0.014704
Previous Claims	0.001869	0.043065	-0.004123	0.001989	1.000000	-0.001172	0.036816	0.003001	0.046874
Vehicle Age	-0.002455	-0.000500	0.001232	0.000316	-0.001172	1.000000	0.000508	0.003132	0.000391
Credit Score	0.002842	-0.201423	-0.001714	0.012016	0.036816	0.000508	1.000000	0.000493	-0.026014
Insurance Duration	-0.000064	0.000379	-0.000265	0.002487	0.003001	0.003132	0.000493	1.000000	-0.000028
Premium Amount	-0.002430	-0.012390	-0.000976	0.014704	0.046874	0.000391	-0.026014	-0.000028	1.000000

تصویر ۳: کوواریانس ویژگی‌ها

همان‌طور که از اعداد کوواریانس مشاهده می‌کنیم، هنوز ارتباط چندان مشخصی بین ویژگی‌ها یا ویژگی هدف نمی‌یابیم. اگرچه حذف صفات نامرتبط ممکن است کار درستی به نظر آید، اما باید توجه داشته باشیم که کوواریانس، تنها مقدار ارتباط خطی بین ویژگی‌ها را نشان می‌دهد.

علاوه بر این، به عنوان انسان، ممکن است از روی تصاویر و نمودارها نتوانیم برخی از ارتباطات را به‌خوبی شناسایی کنیم.

پیش‌پردازش

در مرحله پیش‌پردازش داده‌ها، ابتدا مقادیر از دست رفته را پر کرده‌ایم. برای صفات عددی، از مقدار میانگین آن صفت و برای صفات غیرعددی، از پرتکرارترین مقدار آن‌ها استفاده کرده‌ایم.

در ادامه، برای یادگیری بهتر، تمام مقادیر غیرعددی را به صورت one-hot encoding تبدیل کرده‌ایم و برای مقادیر عددی نیز تمامی آن‌ها را standardize کرده‌ایم. لازم به ذکر است که برای دو صفتی که قبلاً ذکر کردیم، ابتدا لگاریتم اعمال کرده و سپس استانداردسازی را انجام داده‌ایم.

نکته قابل توجه دیگر این است که این پیش‌پردازش‌ها باید برای نمونه‌های تست نیز انجام شود تا مدل آموزش‌دیده بتواند به درستی روی آن‌ها پیش‌بینی انجام دهد.

آموزش و ارزیابی

برای ارزیابی مدل‌هایی که قرار است آموزش دهیم، از معیار ارزیابی RMSLE[1] استفاده کرده‌ایم.

فرمول محاسبه RMSLE به صورت زیر است:

$$RMSLE = \sqrt{\Sigma[(\log(y + 1) - \log(\hat{y} + 1))^2]/n}$$

استفاده از RMSLE به ویژه در مواردی که داده‌ها دارای مقادیر بزرگ و کوچک هستند(همانند Annual Income و Premium Amount)، مفید است و به ما کمک می‌کند تا تأثیر مقادیر بزرگ را کاهش دهیم و دقت پیش‌بینی‌های مدل را بهبود بخشیم.

رگرسیون خطی

اولین مدلی که آموزش داده‌ایم، مدل رگرسیون خطی ساده می‌باشد. این مدل با استفاده از 3-fold cross-validation آموزش داده شده است. با توجه به معیاری که مشخص کرده‌ایم، در هر مرحله یادگیری، دقت زیر را داشته است:

[-0.16493586, -0.16511276, -0.16443987]

استفاده از 3-fold cross-validation به ما این امکان را می‌دهد که مدل را بر روی بخش‌های مختلف داده‌ها آزمایش کنیم و از این طریق ارزیابی بهتری از عملکرد آن به دست آوریم. این روش به ما

1Root Mean Squared Logarithmic Error

2 با توجه به این معیار، هر چه مقدار RMSLE به صفر نزدیک‌تر باشد، نشان‌دهنده دقت بالاتر مدل است

کمک می‌کند تا از overfitting جلوگیری کنیم و اطمینان حاصل کنیم که مدل به خوبی بر روی داده‌های جدید عمل می‌کند.

درخت تصمیم رگرسیونی

مدل بعدی که آموزش داده‌ایم، رگرسیون با استفاده از درخت تصمیم بوده است. این مدل نیز با استفاده از 3-fold cross-validation و Randomized Search بهترین پارامترها را برای آن مشخص کرده‌ایم. مقادیر پارامترهای انتخاب شده را می‌توانید در زیر مشاهده کنید:

{'min_samples_split': 16, 'min_samples_leaf': 10, 'max_features': 'sqrt', 'max_depth': 5}

همچنین، دقت این مدل کمی بهتر از مدل قبلی شده است و برابر با مقدار زیر می‌باشد:

0.1642220362668037